光碟資料庫中主題分佈頻率的初步調查

吳政叡 (Cheng-Juei Wu)

輔仁大學圖書資訊系教授

E-mail: lins1022@mails.fju.edu.tw

中文摘要

先前作者藉由在國家圖書館NBINET上進行的幾個主題檢索個案分析發現，即使透過主題複分的方式，資料分佈極為不平均的情形，仍然造成使用者資訊超載的負荷。本文乃接續先前的研究，以圖書資訊界經常使用的library Literature和Lisa兩種光碟資料庫為對象，調查是否也有一如NBINET的主題分佈極為不平均現象，因而造成使用者的資訊超載。初步調查結果顯示，光碟資料庫上主題分佈不平均現象亦存在，並足以造成使用者的資訊負荷超載。同時類似Lisa這種純粹以期刊文章中關鍵詞為基礎的主題處理方式，其主題分佈頻率有更高度集中的傾向，因此也會對使用者造成更嚴重的資訊超載問題。

==========

A Primary Investigation of Subject Frequency in Library Literature and Lisa

In my previous study based on NBINET, which is the biggest bibliographic database in Taiwan, I discovered that the current practice of subject subdivision alone was not sufficient to avoid information overload. Using the same methodology, I investigate the situation of subject subdivision on two databases, Library Literature and Lisa, which are very frequently used databases in Library and Information Science in this work. The experimental result indicates that the same uneven distribution of subject subdivision is also existed on databases. Subsequently, this posed a very severe information overload problem

for users. In addition, it seems that the information overload problem of Lisa is severer than that of Library Literature.

關鍵字：主題分析，主題分怖，主題複分，標題，資訊超載，模糊邏輯，Subject Analysis，Subject Frequency，Subject Subdivision，Information Overload，Lisa，Library Literature，Fuzzy Logic。

一、前言

人類認知與處理能力的有限，及其所產生的資訊超載現象，是自遠古以來即存在的事實。當資料累積超過一定數量時，單憑人類的記憶能力無法應付時，如何對資料加以適當描述，成為有效利用和整理資料的必然作法，因此自古以來圖書著錄一直是圖書館員的首要核心工作。

除了基本的書目資料，如書名、作者、出版社、與出版日期等，分類號和主題（subject）可以說是圖書著錄人員用來揭露資料內容的兩個主要工具。其中分類號因為肩負排架功能，祇能有一個，因此分類號不若主題來的靈活，所以主題可以說是目前著錄人員用來揭露資料內容最重要的工具。事實上，根據研究顯示，主題檢索可說是圖書館使用者在線上檢索時最常用的方法。[註1-2]

主題（subject）或標題（subject heading）在黃淵泉《中文圖書分類編目學》一書定義為：「一個詞或一組字用以表示資料討論的主題。」 [註 3] 標題的重要性由以前圖書館在卡片目錄盛行時代，一般均會有書名目錄、分類目錄、作者目錄、與標題目錄等四種目錄可以窺知。[註4]

以往圖書館界對於主題或標題的探討，大都集中在控制詞彙的應用（即實踐主題標目的統一原則[註5]）與詞彙的選用、主題標目的結構[註6]、主題標目的複分方式[註7]、主題分析的作法[註8-9]、標題表缺失 [註 10-11]等，近年來雖然也有許多探討主題與檢索效益的關係，例如『中文標題檢索效益之研究--以國立臺灣大學TULIPS系統為例』[註12]，不過仍然是在傳統的主題標目結構下來探討。

在「模糊邏輯在主題分析的應用：標題權值的計算方式」一文中[註 13]，作者因為利用國家圖書館的全國圖書書目資訊網（簡稱NBINET）來收集書目資訊時，親身體驗到資訊超載的經驗，乃針對主題分佈頻率不平均的情況做初步的調查，結果發現圖書館界所使用的主題複分作法，並沒有解決分佈頻率極為不平均的現象，從而造成使用者的資訊超載。在該文中作者提出利用模糊邏輯來解決此問題的建議，並且提出一套主題權值的計算公式。

本文乃延續前面的研究，針對期刊文章的主題分佈頻率情況做初步調查，來了解期刊文章是否也有一如書本主題分佈極為不平均的現象。此次初步調查選定的對象，為圖書資訊界經常使用的Library Literature和Lisa兩種光碟資料庫，其主題的製作方式，恰好是兩種具代表性但截然不同的典型。Library Literature採用類似書目資料主題複分的手法，對所收集的期刊文章予以加工，來產生資料庫中的subject欄位；Lisa資料庫中subject欄位產生的方式，似乎是純粹以期刊文章中作者所給的關鍵詞為基礎，來產生資料庫中的subject欄位。

二、Library Literature主題檢索個案分析

由於Library Literature主要是收藏圖書與資訊相關領域的期刊文章，因此作者選擇「cataloging」與「computer」這二個核心和熟悉的詞彙，來查看Library Literature中它們主題複分的情形。

就「cataloging」而言，總共有2110筆資料分佈在102個以「cataloging」起頭的各式各樣複分標題 [註 14]，例如：CATALOGING、CATALOGING -ADMINISTRATION、CATALOGING -ANALYTIC-ENTRY等。每一標題平均有20.68筆資料，表面看來還不會造成使用者的資訊超載，似乎利用複分可以有效來解決資料數量過多的問題。然而進一步研究卻顯示標題間的頻率分佈情況是極為不平均。

以此個案「cataloging」為例，有最多資料筆數的前四個標題分別為CATALOGING -AUTOMATION、CATALOGING -COOPERATIVE、CATALOGING -ADMINISTRATION、CATALOGING -TEACHING，分別有593、234、123、91筆資料。因此前四多資料的標題總共有1041筆，佔全體資料筆數的49.3% （幾近一半），同時不超過5筆資料的標題計有59個，佔全體複分標題的57.8%（超過一半以上）。

此種資料高度集中於極少數複分標題的現象，固然是忠實反映了此時期的學術研究焦點與興趣；不過就使用者和資料利用的角度而言，卻是個棘手的問題。因為這意謂著越是熱門和使用者感興趣的主題，其資訊超載的情況可能越是嚴重。

從以上的個案分析，也顯示光是依靠傳統主題複分的方式，並不足以解決資訊超載的問題，因為Library Literature是採用類似書目資料主題複分的手法，對所收集的期刊文章予以加工，來產生資料庫中的subject欄位。以「cataloging」開頭的複分標題來說，前四多資料筆數的標題都有近百筆（或以上）記錄，顯然是

超過使用者的負荷能力。

作者另外選擇較偏資訊領域的「computer」標題，來查看其在Library Literature中主題複分的情形。以「computer」來說，總共有3994筆資料分佈在116個以「computer」開頭的複分標題 [註 15]，例如：COMPUTER-ASSISTED-INSTRUCTION和COMPUTER-BULLETIN-BOARDS等，因此每一標題平均有34.43筆資料。由總資料筆數的差異（3994對比2110），似乎可以看出，整體而言，主題「computer」較主題「cataloging」更為熱門許多。

以此個案「computer」為例，有最多資料筆數的前四個標題分別為COMPUTER-ASSISTED-INSTRUCTION、COMPUTER-SOFTWARE、COMPUTER-SOFTWARE-REVIEWS、COMPUTER-SOFTWARE-EVALUATION，分別有695、462、393、325筆資料。因此前四多資料的標題總共有1875筆，佔全體資料筆數的47.0% （幾近一半），同時不超過5筆資料的標題計有67個，佔全體複分標題的57.8%（超過一半以上）。

由上述兩個案「cataloging」和「computer」的統計數據對照，顯示標題間頻率分佈極為不平均的情況是一致的。事實上，較熱門的標題（如「computer」）可能會使讀者有更嚴重資訊超載現象，這可從下列的數據窺知：

（1）標題「computer」中，資料筆數第五多的標題尚有313筆資料；而標題「cataloging」資料筆數第四多的標題祇剩91筆資料。

（2）標題「computer」中，資料筆數前九多的標題，其筆數都超過100，直到第十多的標題才降至95筆資料。

（3） 標題「computer」中，每一標題平均有34.43筆資料；而標題「cataloging」的平均資料筆數是20.68筆資料。

三、Lisa主題檢索個案分析

由於Lisa 一如Library Literature般，主要是收藏圖書與資訊相關領域的期刊文章，不過其資料庫中subject欄位產生的方式，似乎是直接選用期刊文章中作者所給的關鍵詞，並未如Library Literature般使用圖書館書目資料的主題複分手法和形式。為了對比，此處仍以「cataloging」與「computer」這二個詞彙，來查看它們在Lisa中主題分佈頻率的情形。

就「cataloging」而言，由於拼字差異分為cataloging和cataloguing兩種[註 16]，其中又以cataloguing較為普遍，因此這裏以cataloguing來進行下面的討論。總共有

6315筆資料分佈在26個以「cataloguing」起頭的主題 [註 17]，例如：CATALOGUING、CATALOGUING AND INDEXING GROUP等。每一主題平均有242.88筆資料，顯示純粹以期刊文章的關鍵詞為基礎的主題，其主題複分的數量更少，使得資料更形集中，更容易造成使用者的資訊超載。

以「cataloguing」來說，有最多資料筆數的前四個主題分別為CATALOGUING、CATALOGUING RULES、CATALOGUING IN PUBLICATION、CATALOGUING AIDS，分別有6029、134、91、17筆資料。可以看出資料筆數的高度集中，光是CATALOGUING一個主題的資料筆數，就佔全體資料筆數的95.5% （幾近全部）。這顯示未經加工的關鍵詞，其主題分佈頻率有非常高度集中的傾向，連帶造成更嚴重資訊超載的情形。

以偏資訊領域的「computer」來說，則總共有5820筆資料分佈在174個以「computer」開頭的主題 [註 18]，例如：COMPUTER AIDED DESIGN和COMPUTER-HUMAN INTERACTION等，因此每一主題平均有33.49筆資料。

以此個案「computer」為例，有最多資料筆數的前四個主題分別為COMPUTER APPLICATIONS、COMPUTER ASSISTED INSTRUCTION、COMPUTER SCIENCE、COMPUTER SECURITY，分別有2753、1101、474、350筆資料。因此前四多資料的主題總共有4678筆，佔全體資料筆數的80.4% ，同時不超過5筆資料的主題計有151個，佔全體主題數量的86.8%。

由上述的個案分析可知，不管是「cataloguing」或「computer」，純粹以期刊文章中所給關鍵詞為基礎的主題處理方式，其主題分佈頻率有非常高度集中的傾向，這造成對使用者更嚴重的資訊超載問題。

四、結語

根據研究顯示，主題檢索是使用者在線上檢索時最常使用的方法，因此其重要性是不言而喻的。但是先前作者藉由在國家圖書館NBINET上進行的幾個主題檢索個案分析發現，即使透過主題複分的方式，主題分佈頻率極為不平均的情形，仍然造成使用者資訊超載的負荷。本文乃接續先前的研究，以圖書資訊界經常使用的Library Literature和Lisa兩種光碟資料庫為對象，調查是否也有一如NBINET的主題分佈頻率極為不平均現象，因而造成使用者的資訊超載。

經過作者在Library Literature和Lisa兩種光碟資料庫上，針對「cataloging」與「computer」所做的初步調查結果顯示[註 19]，光碟資料庫上主題分佈頻率不平均的現象亦同樣存在，並且足以造成使用者的資訊負荷超載。

另一方面，由於Library Literature和Lisa兩種光碟資料庫，其主題的製作方式，恰好是兩種截然不同的典型。Library Literature採用類似書目資料主題複分的手法，來產生資料庫中的subject欄位；相反的，Lisa似乎是純粹以期刊文章中作者所給的關鍵詞為基礎，來產生資料庫中的subject欄位。由兩者統計數據的對照中，作者也初步發現，類似Lisa這種主題處理方式，其主題分佈頻率有更高度集中的傾向，因此也會對使用者造成更嚴重的資訊超載問題。

至於針對此種期刊文章中主題分佈頻率不平均現象的可能解決方案，作者已在「模糊邏輯在主題分析的應用：標題權值的計算方式」一文中，提出以模糊邏輯為對策的方案，有關模糊邏輯和該解決方案的詳細闡釋，請讀者參閱該文。[註 20]

註釋

註 1：盧秀菊，「中文主題標目與標題表」，中國圖書館學會會報59 期（民 86 年 12 月），頁26。

註 2：曾繁絹，「中文標題檢索效益之研究--以國立臺灣大學TULIPS系統為例」，大學圖書館第2卷第1 期（民 87 年 1 月），頁101。

註 3：黃淵泉，中文圖書分類編目學（台北市：學生書局，民 85 年 4 月），頁 11。

註 4：同註3，頁27-28。

註 5：陳麥麟屏和林國強，美國國會圖書館主題標目（台北市：三民書局，民 78 年 12 月），頁27-29。

註 6：同註5，頁41-48。

註 7：同註5，頁49-73。

註 8：陳昭珍，「主題檢索理論之探討－主題分析（上）」，書農9 期（民 81 年 12 月），頁 11-27。

註 9：陳佳君，「從知識結構探討主題分析」，書府16 期（民 84 年 6 月），頁 38-47。

註10：同註1，頁 31-38。

註11：侯漢清，「評《中文圖書標題表》—兼談標題表的敘詞化改進」，圖書與資訊學刊31 期（民 88 年 11 月），頁 17-23。

註12：同註2，頁100-123。

註13：吳政叡，「模糊邏輯在主題分析的應用：標題權值的計算方式」，圖書與資訊學刊40 期（民 91 年 2 月），頁 10-17。

註14：實際檢索日期是2001年12月26日，使用輔仁大學圖書館所有的Library Literature光碟資料庫。

註15：檢索日期與場所相同於註14。

註16：Lisa中cataloging和cataloguing兩種主題均有，而且是分別排列。相對於總共有6315筆資料分佈在26個以「cataloguing」起頭的主題；cataloging開頭的主題，祇有41筆資料分佈在19項。

註17：實際檢索日期是2001年12月26日，使用輔仁大學圖書館所有的Lisa光碟資料庫。

註18：檢索日期與場所相同於註17。

註19：由於拼字差異，在Lisa上是使用cataloguing而非cataloging。

註20：同註13，頁13-14。